Conformal calibration guarantees for reliable predictions

Johanna Ziegel

joint work with

Sam Allen, Georgios Gavrilopoulos, Alexander Henzi, Gian-Reto Kleger

ETH Zurich

Insurance Data Science Conference
Bayes Business School, London

19–20 June, 2025

Introduction

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathcal{Y} = \mathbb{R}$)
  - Claim size. ($Y \in \mathcal{Y} = [0, \infty)$)
  - Event of rain tomorrow in London. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Default of credit card client. ($Y \in \mathcal{Y} = \{0, 1\}$)
- Point prediction for Y:
  - Single valued "best guess" $Z \in \mathcal{Y}$.
  - Does not quantify uncertainty, but maybe useful/necessary e.g. for pricing.
  - If X is information available for prediction, often, we try to approximate $\mathbb{E}[Y \mid X]$.
- Probabilistic prediction for Y:
  - Quantify uncertainty of Y by specifying a distribution F on $\mathcal{Y}$.
  - If X is information available for prediction, F should approximate $\mathcal{L}(Y \mid X)$.
  - Other possibilities to quantify uncertainty of Y: prediction intervals, predictions of some measure of variability, ...

Introduction

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathcal{Y} = \mathbb{R}$)
  - Claim size. ($Y \in \mathcal{Y} = [0, \infty)$)
  - Event of rain tomorrow in London. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Default of credit card client. ($Y \in \mathcal{Y} = \{0, 1\}$)
- Point prediction for Y:
  - Single valued "best guess" $Z \in \mathcal{Y}$.
  - Does not quantify uncertainty, but maybe useful/necessary e.g. for pricing.
  - If X is information available for prediction, often, we try to approximate $\mathbb{E}[Y \mid X]$.
- Probabilistic prediction for Y:
  - Quantify uncertainty of Y by specifying a distribution F on $\mathcal{Y}$.
  - If X is information available for prediction, F should approximate $\mathcal{L}(Y \mid X)$.
  - Other possibilities to quantify uncertainty of Y: prediction intervals, predictions of some measure of variability, ...

Introduction

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathcal{Y} = \mathbb{R}$)
  - Claim size. ($Y \in \mathcal{Y} = [0, \infty)$)
  - Event of rain tomorrow in London. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Default of credit card client. ($Y \in \mathcal{Y} = \{0, 1\}$)
- Point prediction for Y:
  - Single valued "best guess" $Z \in \mathcal{Y}$.
  - Does not quantify uncertainty, but maybe useful/necessary e.g. for pricing.
  - If X is information available for prediction, often, we try to approximate $\mathbb{E}[Y \mid X]$.
- Probabilistic prediction for Y:
  - Quantify uncertainty of Y by specifying a distribution F on $\mathcal{Y}$.
  - If X is information available for prediction, F should approximate $\mathcal{L}(Y \mid X)$.
  - Other possibilities to quantify uncertainty of Y: prediction intervals, predictions of some measure of variability, ...

Introduction

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathcal{Y} = \mathbb{R}$)
  - Claim size. ($Y \in \mathcal{Y} = [0, \infty)$)
  - Event of rain tomorrow in London. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Default of credit card client. ($Y \in \mathcal{Y} = \{0, 1\}$)
- Point prediction for Y:
  - Single valued "best guess" $Z \in \mathcal{Y}$.
  - Does not quantify uncertainty, but maybe useful/necessary e.g. for pricing.
  - If X is information available for prediction, often, we try to approximate $\mathbb{E}[Y \mid X]$.
- Probabilistic prediction for Y:
  - Quantify uncertainty of Y by specifying a distribution F on $\mathcal{Y}$.
  - If X is information available for prediction, F should approximate $\mathcal{L}(Y \mid X)$.
  - Other possibilities to quantify uncertainty of Y: prediction intervals, predictions of some measure of variability, ...

Introduction

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathcal{Y} = \mathbb{R}$)
  - Claim size. ($Y \in \mathcal{Y} = [0, \infty)$)
  - Event of rain tomorrow in London. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Default of credit card client. ($Y \in \mathcal{Y} = \{0, 1\}$)
- Point prediction for Y:
  - Single valued "best guess" $Z \in \mathcal{Y}$.
  - Does not quantify uncertainty, but maybe useful/necessary e.g. for pricing.
  - If X is information available for prediction, often, we try to approximate $\mathbb{E}[Y \mid X]$.
- Probabilistic prediction for Y:
  - Quantify uncertainty of Y by specifying a distribution F on $\mathcal{Y}$.
  - If X is information available for prediction, F should approximate $\mathcal{L}(Y \mid X)$.
  - Other possibilities to quantify uncertainty of Y: prediction intervals, predictions of some measure of variability, ...

Quality criteria for predictions

- What is *calibration* of predictions?
- How do we calibrate predictions?
- How do we compare predictions and how is related to calibration?

- Forecasts are usually sequential but many concepts are easier to understand in a "hypothetical" one-period setting.
- Future outcome Y and forecasts Z or F are both random and defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Quality criteria for predictions

- What is *calibration* of predictions?
- How do we calibrate predictions?
- How do we compare predictions and how is related to calibration?

- Forecasts are usually sequential but many concepts are easier to understand in a "hypothetical" one-period setting.
- Future outcome Y and forecasts Z or F are both random and defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Quality criteria for predictions

- What is *calibration* of predictions?
- How do we calibrate predictions?
- How do we compare predictions and how is related to calibration?

- Forecasts are usually sequential but many concepts are easier to understand in a "hypothetical" one-period setting.
- Future outcome Y and forecasts Z or F are both random and defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Quality criteria for predictions

- What is *calibration* of predictions?
- How do we calibrate predictions?
- How do we compare predictions and how is related to calibration?

- Forecasts are usually sequential but many concepts are easier to understand in a "hypothetical" one-period setting.
- Future outcome Y and forecasts Z or F are both random and defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Simplest case: Binary outcomes

- $Y \in \{0, 1\}$
  - Event of rain tomorrow in London. ($Y \in \{0, 1\}$)
  - Default of credit card client. ($Y \in \{0, 1\}$)
- Distribution of Y is characterised by probability of $\{Y = 1\}$:
  Probabilistic prediction is random variable $p \in [0, 1]$.
- Since $\mathbb{P}(Y = 1 \mid X) = \mathbb{E}(Y \mid X)$,
  - p is a prediction for the conditional distribution of Y;
  - p is a prediction for the conditional mean of Y.

Definition

A probability prediction $p \in [0, 1]$ for $Y \in \{0, 1\}$ is *calibrated* (or *reliable*) if

$$\mathbb{E}[Y \mid p] = \mathbb{P}(Y = 1 \mid p) = p.$$

Predicted probabilities should align with observed frequencies.

Simplest case: Binary outcomes

- $Y \in \{0, 1\}$
  - Event of rain tomorrow in London. ($Y \in \{0, 1\}$)
  - Default of credit card client. ($Y \in \{0, 1\}$)
- Distribution of Y is characterised by probability of $\{Y = 1\}$: Probabilistic prediction is random variable $p \in [0, 1]$.
- Since $\mathbb{P}(Y = 1 \mid X) = \mathbb{E}(Y \mid X)$,
  - p is a prediction for the conditional distribution of Y;
  - p is a prediction for the conditional mean of Y.

Definition

A probability prediction $p \in [0, 1]$ for $Y \in \{0, 1\}$ is *calibrated* (or *reliable*) if

$$\mathbb{E}[Y \mid p] = \mathbb{P}(Y = 1 \mid p) = p.$$

Predicted probabilities should align with observed frequencies.

Simplest case: Binary outcomes

- $Y \in \{0, 1\}$
  - Event of rain tomorrow in London. ($Y \in \{0, 1\}$)
  - Default of credit card client. ($Y \in \{0, 1\}$)
- Distribution of Y is characterised by probability of $\{Y = 1\}$: Probabilistic prediction is random variable $p \in [0, 1]$.
- Since $\mathbb{P}(Y = 1 \mid X) = \mathbb{E}(Y \mid X)$,
  - p is a prediction for the conditional distribution of Y;
  - p is a prediction for the conditional mean of Y.

Definition

A probability prediction $p \in [0, 1]$ for $Y \in \{0, 1\}$ is *calibrated* (or *reliable*) if

$$\mathbb{E}[Y \mid p] = \mathbb{P}(Y = 1 \mid p) = p.$$

Predicted probabilities should align with observed frequencies.

Simplest case: Binary outcomes

- $Y \in \{0,1\}$
  - Event of rain tomorrow in London. ($Y \in \{0,1\}$)
  - Default of credit card client. ($Y \in \{0,1\}$)
- Distribution of Y is characterised by probability of $\{Y = 1\}$: Probabilistic prediction is random variable $p \in [0,1]$.
- Since $\mathbb{P}(Y = 1 \mid X) = \mathbb{E}(Y \mid X)$,
  - p is a prediction for the conditional distribution of Y;
  - p is a prediction for the conditional mean of Y.

Definition

A probability prediction $p \in [0,1]$ for $Y \in \{0,1\}$ is *calibrated* (or *reliable*) if

$$\mathbb{E}[Y \mid p] = \mathbb{P}(Y = 1 \mid p) = p.$$

Predicted probabilities should align with observed frequencies.

Example

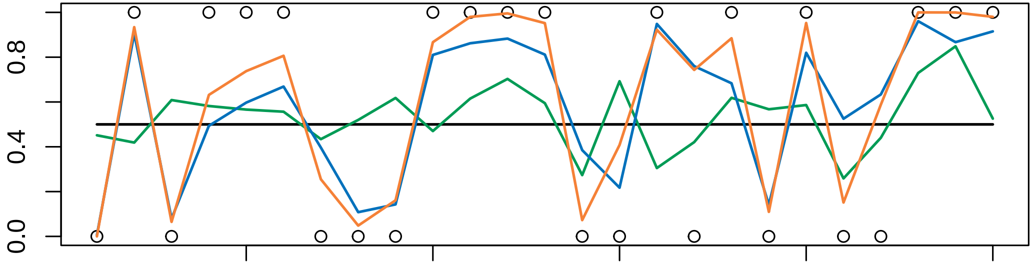
Let $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$ be independent, and

$$\mathbb{P}(Y = 1 \mid X_1, X_2) = \Phi(X_1 + X_2).$$

Predictions:

$$p_0 = 1/2, \quad p_1 = \Phi\left(\frac{X_1}{\sqrt{3}}\right), \quad p_2 = \Phi\left(\frac{X_2}{\sqrt{2}}\right), \quad p_3 = \Phi(X_1 + X_2).$$

- All predictions are calibrated.

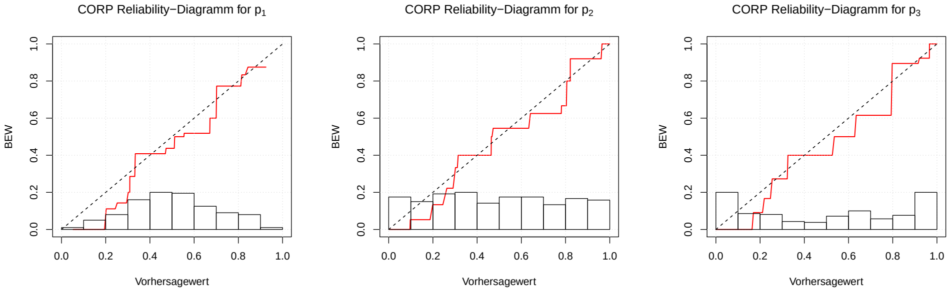


Diagnostics to assess calibration: Reliability diagrams

Data: $(p^1, Y_1), \ldots, (p^n, Y_n)$

Simulation example

$X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$ independent, $\mathbb{P}(Y = 1 \mid X_1, X_2) = \Phi(X_1 + X_2)$,
$p_1 = \Phi(X_1/\sqrt{3})$, $p_2 = \Phi\left(X_2/\sqrt{2}\right)$, $p_3 = \Phi(X_1 + X_2)$, $n = 200$.



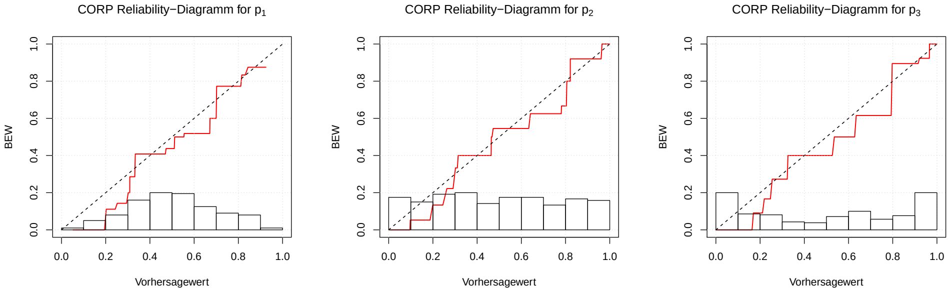
(Dimitriadis et al., 2021)

Diagnostics to assess calibration: Reliability diagrams

Data: $(p^1, Y_1), \dots, (p^n, Y_n)$

Simulation example

$X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$ independent, $\mathbb{P}(Y = 1 \mid X_1, X_2) = \Phi(X_1 + X_2)$,
$p_1 = \Phi(X_1/\sqrt{3})$, $p_2 = \Phi\Big(X_2/\sqrt{2}\Big)$, $p_3 = \Phi(X_1 + X_2)$, $n = 200$.



(Dimitriadis et al., 2021)

Real-valued outcomes: Probabilistic predictions

- $Y \in \mathbb{R}$.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathbb{R}$)
  - Claim size. ($Y \in \mathcal{Y} = [0, \infty)$)
- Quantify uncertainty of Y by a probabilistic prediction F.
  - F is a distribution on $\mathbb{R}$ (typically specified as a CDF).
- If X is information available for prediction, F should approximate $\mathcal{L}(Y \mid X)$.

Real-valued outcomes: Probabilistic predictions

- $Y \in \mathbb{R}$.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathbb{R}$)
  - Claim size. ($Y \in \mathcal{Y} = [0, \infty)$)
- Quantify uncertainty of Y by a probabilistic prediction F.
  - F is a distribution on $\mathbb{R}$ (typically specified as a CDF).
- If X is information available for prediction, F should approximate $\mathcal{L}(Y \mid X)$.

Illustration: Point and probabilistic predictions

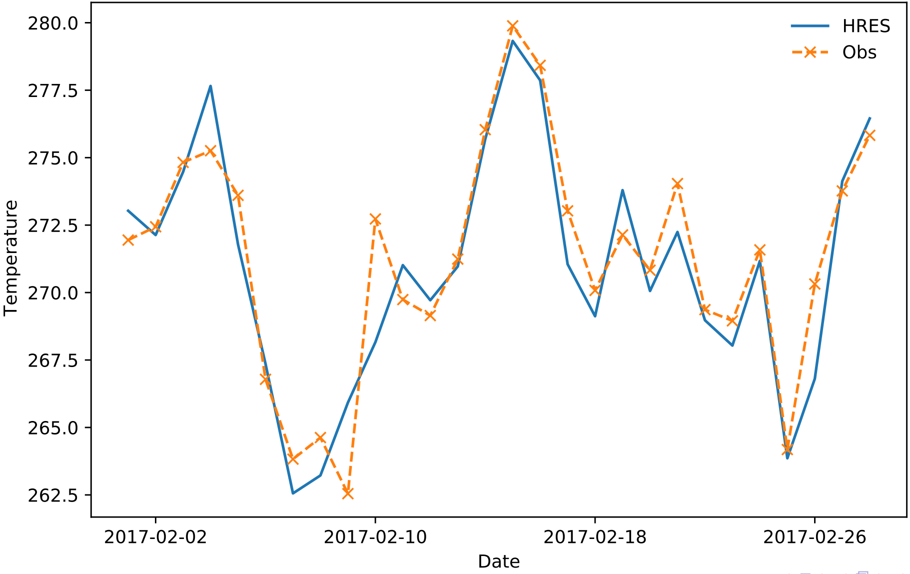
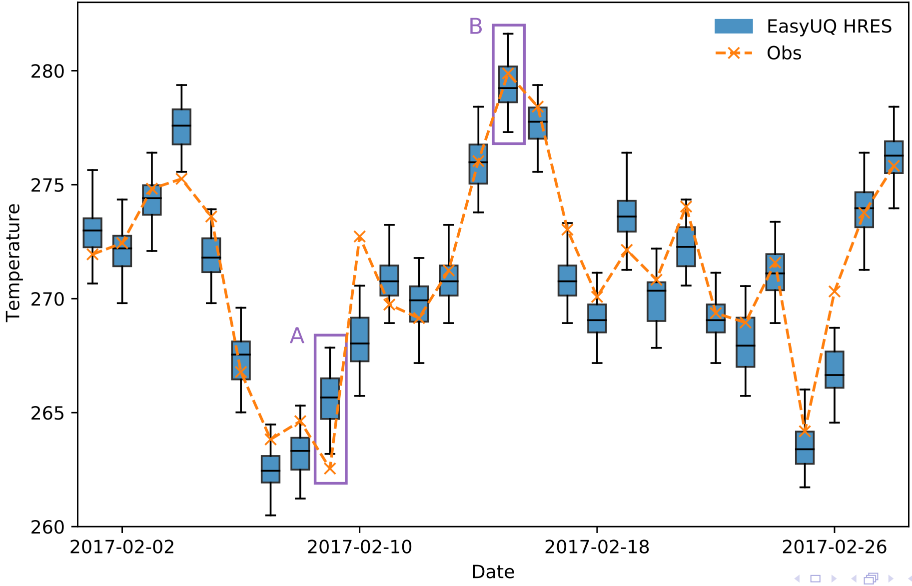


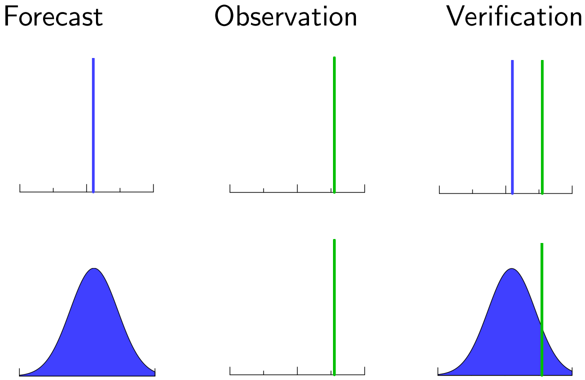
Illustration: Point and probabilistic predictions



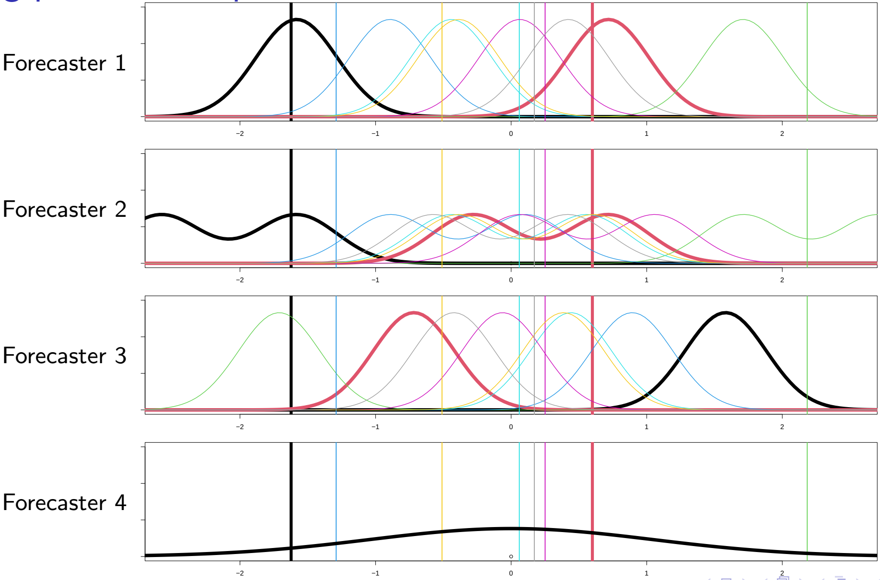
Probabilistic and point predictions

“Tomorrow at 12:00
temperature will be 17.5°C.”

“Tomorrow at 12:00
temperature will be $\mathcal{N}(17.5, \sigma^2)$.”



Evaluating probabilistic predictions



Calibration: Compatibility between forecasts and observations

Probabilities derived from predictive distributions should align with observed frequencies.

Most popular: Probabilistic calibration/"Flat PIT histogram"

$$F_i(Y_i) \sim \mathrm{UNIF}(0,1) \quad \text{for all } i$$

- $Y_i \in \mathbb{R}$, F_i predictive CDF for Y_i
- Suitable randomization if F_i is not continuous
- Closely related to validity of conformal predictive systems. Ensures marginal coverage of prediction intervals.
- **Binary outcomes**: $Y_i \in \{0,1\} : \mathbb{P}(Y_i = 1 | p_i) = p_i$
- *Many* notions of calibration, except for binary outcomes...

Calibration: Compatibility between forecasts and observations

Probabilities derived from predictive distributions should align with observed frequencies.

Most popular: Probabilistic calibration/"Flat PIT histogram"

$$F_i(Y_i) \sim \mathrm{UNIF}(0,1) \quad \text{for all } i$$

- $Y_i \in \mathbb{R}$, F_i predictive CDF for Y_i
- Suitable randomization if F_i is not continuous
- Closely related to validity of conformal predictive systems. Ensures marginal coverage of prediction intervals.
- **Binary outcomes**: $Y_i \in \{0,1\} : \mathbb{P}(Y_i = 1 | p_i) = p_i$
- *Many* notions of calibration, except for binary outcomes...

Calibration: Compatibility between forecasts and observations

Probabilities derived from predictive distributions should align with observed frequencies.

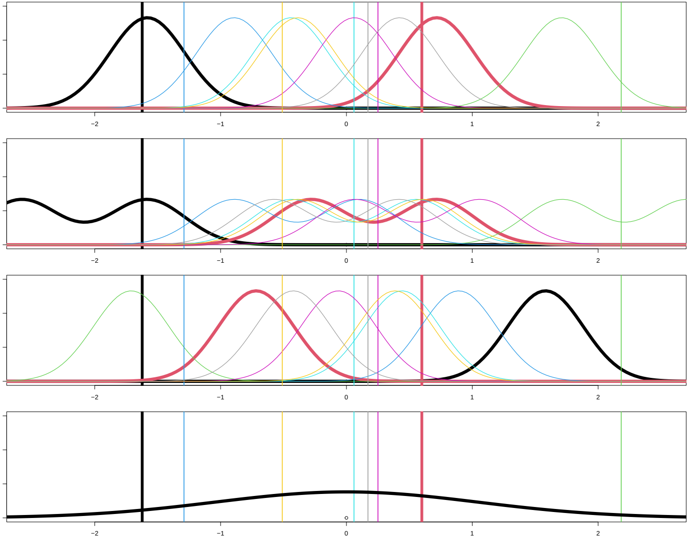
Most popular: Probabilistic calibration/"Flat PIT histogram"

$$F_i(Y_i) \sim \mathrm{UNIF}(0,1) \quad \text{for all } i$$

- $Y_i \in \mathbb{R}$, F_i predictive CDF for Y_i
- Suitable randomization if F_i is not continuous
- Closely related to validity of conformal predictive systems. Ensures marginal coverage of prediction intervals.
- **Binary outcomes**: $Y_i \in \{0,1\} : \mathbb{P}(Y_i = 1|p_i) = p_i$
- *Many* notions of calibration, except for binary outcomes...

Evaluating probabilistic predictions

$$\mu \sim \mathcal{N}(0,1), \quad Y \sim \mathcal{N}(\mu, 0.09)$$



Probabilistic calibration ✓

Probabilistic calibration ✓

Probabilistic calibration ×

Probabilistic calibration ✓

Many notions of calibration ...

Auto-calibration:
$\mathbb{P}(Y_i > y \mid F_i) = 1 - F_i(y)\ \forall y$
$\mathcal{L}(Y_i \mid F_i) = F_i$

$\Downarrow$

Isotonic calibration:
$\mathbb{P}(Y_i > y \mid \mathcal{A}(F_i)) = 1 - F_i(y)\ \forall y$
$\mathcal{L}(Y_i \mid \mathcal{A}(F_i)) = F_i$

$\swarrow \quad \searrow$

Threshold calibration:
$\mathbb{P}(Y_i > y \mid F_i(y)) = 1 - F_i(y)\ \forall y$

$\Downarrow$

Marginal calibration:
$\mathbb{P}(Y_i > y) = 1 - \mathbb{E}F_i(y)\ \forall y$

Quantile calibration:
$q_\alpha(Y_i \mid F_i^{-1}(\alpha)) = F_i^{-1}(\alpha)\ \forall \alpha$

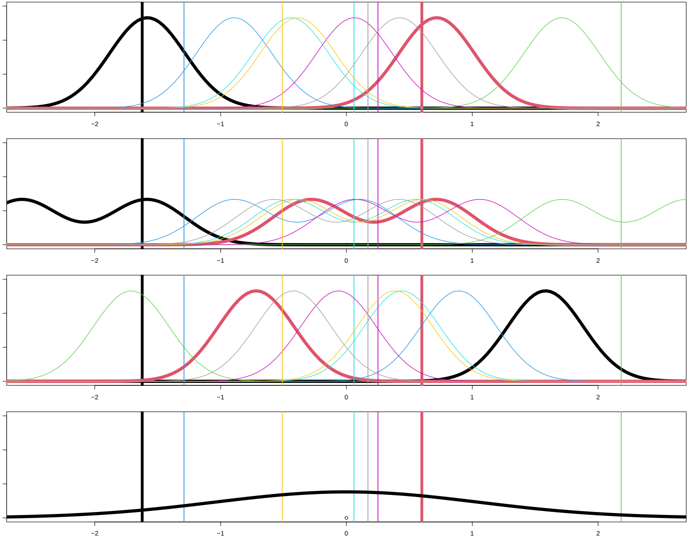
$\Downarrow$

Probabilistic calibration:
$F_i(Y_i) \sim \mathrm{UNIF}(0, 1)$
$\mathbb{P}(F_i(Y_i) < \alpha) \leq \alpha \leq \mathbb{P}(F_i(Y_i-) \leq \alpha)\ \forall \alpha$

And if we want to focus on tails of F_i... (Allen et al., 2025b)

Evaluating probabilistic predictions

$$\mu \sim \mathcal{N}(0,1), \quad Y \sim \mathcal{N}(\mu, 0.09)$$



Auto-calibration ✓
Probabilistic calibration ✓
Marginal calibration ✓

Auto-calibration ×
Probabilistic calibration ✓
Marginal calibration ×

Auto-calibration ×
Probabilistic calibration ×
Marginal calibration ✓

Auto-calibration ✓
Probabilistic calibration ✓
Marginal calibration ✓

- Probabilistic predictions should be calibrated, ideally, *auto-calibrated*.
- Subject to calibration, they should be *sharp* in order to be informative.
- Comparison of probabilistic predictions with proper scoring rules:
  Assign a real-valued score assessing calibration and sharpness simultaneously.

Logarithmic Score (LogS) f density of F

$$\mathrm{LogS}(F, y) = -\log f(y)$$

Continuous Ranked Probability Score (CRPS) F CDF, finite mean

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{y \leq z\})^2 \,\mathrm{d}z$$

Conformal prediction

Goal: Provide predictions with calibration guarantees out-of-sample.

What is at the heart of conformal prediction?

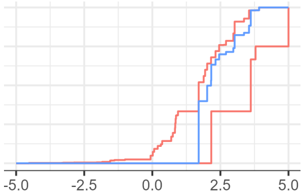
"In-sample calibration yields conformal calibration guarantees."

Predictive system

A set $\Pi \subseteq \mathbb{R} \times [0,1]$ of the form

$$\Pi = \{(y,\tau) \mid \Pi_\ell(y) \leq \tau \leq \Pi_u(y)\}$$

with $\Pi_\ell \leq \Pi_u$ increasing, $\lim_{y\to-\infty} \Pi_\ell(y) = 0$, $\lim_{y\to\infty} \Pi_u(y) = 1$.



Conformal calibration guarantee:

We can construct a predictive system that contains a calibrated CDF.

What is at the heart of conformal prediction?

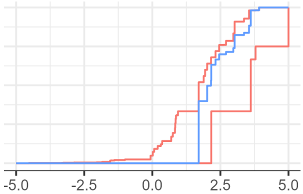
"In-sample calibration yields conformal calibration guarantees."

Predictive system

A set $\Pi \subseteq \mathbb{R} \times [0,1]$ of the form

$$\Pi = \{(y, \tau) \mid \Pi_\ell(y) \leq \tau \leq \Pi_u(y)\}$$

with $\Pi_\ell \leq \Pi_u$ increasing, $\lim_{y \to -\infty} \Pi_\ell(y) = 0$, $\lim_{y \to \infty} \Pi_u(y) = 1$.



Conformal calibration guarantee:

We can construct a predictive system that contains a calibrated CDF.

What is at the heart of conformal prediction?

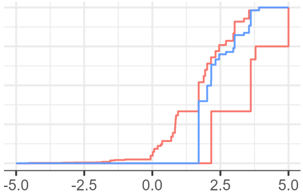
"In-sample calibration yields conformal calibration guarantees."

Predictive system

A set $\Pi \subseteq \mathbb{R} \times [0,1]$ of the form

$$\Pi = \{(y, \tau) \mid \Pi_\ell(y) \leq \tau \leq \Pi_u(y)\}$$

with $\Pi_\ell \leq \Pi_u$ increasing, $\lim_{y \to -\infty} \Pi_\ell(y) = 0$, $\lim_{y \to \infty} \Pi_u(y) = 1$.



Conformal calibration guarantee:

We can construct a predictive system that contains a calibrated CDF.

Example of in-sample calibration:

Let $w_1, \ldots, w_m \in \mathbb{R}$. Define

$$F(y) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{w_i \leq y\}, \quad y \in \mathbb{R}.$$

Draw W uniformly at random from $w_1, \ldots, w_m$.

Then F is *in-sample* probabilistically calibrated, that is,

$$\mathbb{P}(F(W) < \alpha) \leq \alpha \leq \mathbb{P}(F(W-) \leq \alpha), \quad \alpha \in (0, 1).$$

$$F(W) \approx \mathrm{UNIF}(0, 1)$$

Let $W_1, \ldots, W_{n+1} \in \mathbb{R}$ be exchangeable and define for $w \in \mathbb{R}$

$$F^w(y) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\{W_i \leq y\} + \frac{1}{n+1} \mathbb{1}\{w \leq y\}, \quad y \in \mathbb{R},$$

and

$$\Pi_\ell(y) = \inf\{F^w(y) \mid w \in \mathbb{R}\}, \quad \Pi_u(y) = \sup\{F^w(y) \mid w \in \mathbb{R}\},$$

Then,

$$\Pi_\ell(y) \leq F^{W_{n+1}}(y) \leq \Pi_u(y), \quad \text{and}$$

$$\mathbb{P}(F^{W_{n+1}}(W_{n+1}) < \alpha) \leq \alpha \leq \mathbb{P}(F^{W_{n+1}}(W_{n+1}-) \leq \alpha), \quad \alpha \in (0,1).$$

Proof: Conditional on empirical distribution $\hat{\mathbb{P}}_{n+1}$ of $(W_i)_{i=1}^{n+1}$, W_{n+1} is a random draw from $W_1, \ldots, W_{n+1}$. By in-sample probabilistic calibration:

$$\mathbb{P}(F^{W_{n+1}}(W_{n+1}) < \alpha \mid \hat{\mathbb{P}}_{n+1}) \leq \alpha \leq \mathbb{P}(F^{W_{n+1}}(W_{n+1}-) \leq \alpha \mid \hat{\mathbb{P}}_{n+1}) \ldots$$

Let $W_1, \dots, W_{n+1} \in \mathbb{R}$ be exchangeable and define for $w \in \mathbb{R}$

$$F^w(y) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\{W_i \le y\} + \frac{1}{n+1} \mathbb{1}\{w \le y\}, \quad y \in \mathbb{R},$$

and

$$\Pi_\ell(y) = \inf\{F^w(y) \mid w \in \mathbb{R}\}, \quad \Pi_u(y) = \sup\{F^w(y) \mid w \in \mathbb{R}\},$$

Then,

$$\Pi_\ell(y) \le F^{W_{n+1}}(y) \le \Pi_u(y), \quad \text{and}$$

$$\mathbb{P}(F^{W_{n+1}}(W_{n+1}) < \alpha) \le \alpha \le \mathbb{P}(F^{W_{n+1}}(W_{n+1}-) \le \alpha), \quad \alpha \in (0,1).$$

Proof: Conditional on empirical distribution $\hat{\mathbb{P}}_{n+1}$ of $(W_i)_{i=1}^{n+1}$, W_{n+1} is a random draw from $W_1, \dots, W_{n+1}$. By in-sample probabilistic calibration:

$$\mathbb{P}(F^{W_{n+1}}(W_{n+1}) < \alpha \mid \hat{\mathbb{P}}_{n+1}) \le \alpha \le \mathbb{P}(F^{W_{n+1}}(W_{n+1}-) \le \alpha \mid \hat{\mathbb{P}}_{n+1}) \dots$$

(Classical) conformal prediction trick

Use conformity measure $A(\hat{\mathbb{P}}, (x, y))$ to lift the one-dimensional result to general spaces $\mathcal{X} \times \mathcal{Y}$.

Let $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{X} \times \mathbb{R}$ be exchangeable.

- $\hat{\mathbb{P}}^y$: Empirical distribution of $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$ for $y \in \mathbb{R}$
- $\hat{F}^y$: Empirical CDF of

$$W_1 = A(\hat{\mathbb{P}}^y, (X_1, Y_1)), \ldots, W_n = A(\hat{\mathbb{P}}^y, (X_n, Y_n)), w(y) = A(\hat{\mathbb{P}}^y, (X_{n+1}, y))$$

- $\mathbb{P}(F^{Y_{n+1}}(w(Y_{n+1})) < \alpha) \leq \alpha \quad \leq \mathbb{P}(F^{Y_{n+1}}(w(Y_{n+1})-) \leq \alpha)$
- This implies $\mathbb{P}(Y_{n+1} \in C_{n+1}) \geq 1 - \alpha \quad \geq \mathbb{P}(Y_{n+1} \in C_{n+1}^-)$, where

$$C_{n+1} = \{y \in \mathbb{R} \mid F^y(w(y)) \geq \alpha\}.$$

- Predictive CDF available if $y \mapsto F^y(w(y))$, $y \mapsto F^y(w(y)-)$ are increasing.
  (Classical) conformal predictive system

(Classical) conformal prediction trick

Use conformity measure $A(\hat{\mathbb{P}}, (x, y))$ to lift the one-dimensional result to general spaces $\mathcal{X} \times \mathcal{Y}$.

Let $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{X} \times \mathbb{R}$ be exchangeable.

- $\hat{\mathbb{P}}^y$: Empirical distribution of $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$ for $y \in \mathbb{R}$
- $\hat{F}^y$: Empirical CDF of

$$W_1 = A(\hat{\mathbb{P}}^y, (X_1, Y_1)), \ldots, W_n = A(\hat{\mathbb{P}}^y, (X_n, Y_n)), w(y) = A(\hat{\mathbb{P}}^y, (X_{n+1}, y))$$

- $\mathbb{P}(F^{Y_{n+1}}(w(Y_{n+1})) < \alpha) \leq \alpha \quad \leq \mathbb{P}(F^{Y_{n+1}}(w(Y_{n+1})-) \leq \alpha)$
- This implies $\mathbb{P}(Y_{n+1} \in C_{n+1}) \geq 1 - \alpha \quad \geq \mathbb{P}(Y_{n+1} \in C_{n+1}^-)$, where

$$C_{n+1} = \{y \in \mathbb{R} \mid F^y(w(y)) \geq \alpha\}.$$

- Predictive CDF available if $y \mapsto F^y(w(y))$, $y \mapsto F^y(w(y)-)$ are increasing. (Classical) conformal predictive system

(Classical) conformal prediction trick

Use conformity measure $A(\hat{\mathbb{P}}, (x, y))$ to lift the one-dimensional result to general spaces $\mathcal{X} \times \mathcal{Y}$.

Let $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{X} \times \mathbb{R}$ be exchangeable.

- $\hat{\mathbb{P}}^y$: Empirical distribution of $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$ for $y \in \mathbb{R}$
- $\hat{F}^y$: Empirical CDF of

$$W_1 = A(\hat{\mathbb{P}}^y, (X_1, Y_1)), \ldots, W_n = A(\hat{\mathbb{P}}^y, (X_n, Y_n)), w(y) = A(\hat{\mathbb{P}}^y, (X_{n+1}, y))$$

- $\mathbb{P}(F^{Y_{n+1}}(w(Y_{n+1})) < \alpha) \leq \alpha \quad \leq \mathbb{P}(F^{Y_{n+1}}(w(Y_{n+1})-) \leq \alpha)$
- This implies $\mathbb{P}(Y_{n+1} \in C_{n+1}) \geq 1 - \alpha \quad \geq \mathbb{P}(Y_{n+1} \in C_{n+1}^{-})$, where

$$C_{n+1} = \{y \in \mathbb{R} \mid F^y(w(y)) \geq \alpha\}.$$

- Predictive CDF available if $y \mapsto F^y(w(y))$, $y \mapsto F^y(w(y)-)$ are increasing.
  (Classical) conformal predictive system

(Classical) conformal prediction trick

Use conformity measure $A(\hat{\mathbb{P}}, (x, y))$ to lift the one-dimensional result to general spaces $\mathcal{X} \times \mathcal{Y}$.

Let $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \in \mathcal{X} \times \mathbb{R}$ be exchangeable.

- $\hat{\mathbb{P}}^y$: Empirical distribution of $(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, y)$ for $y \in \mathbb{R}$
- $\hat{F}^y$: Empirical CDF of

$$W_1 = A(\hat{\mathbb{P}}^y, (X_1, Y_1)), \dots, W_n = A(\hat{\mathbb{P}}^y, (X_n, Y_n)), w(y) = A(\hat{\mathbb{P}}^y, (X_{n+1}, y))$$

- $\mathbb{P}(F^{Y_{n+1}}(w(Y_{n+1})) < \alpha) \leq \alpha \quad \leq \mathbb{P}(F^{Y_{n+1}}(w(Y_{n+1})-) \leq \alpha)$
- This implies $\mathbb{P}(Y_{n+1} \in C_{n+1}) \geq 1 - \alpha \quad \geq \mathbb{P}(Y_{n+1} \in C_{n+1}^-)$, where

$$C_{n+1} = \{y \in \mathbb{R} \mid F^y(w(y)) \geq \alpha\}.$$

- Predictive CDF available if $y \mapsto F^y(w(y))$, $y \mapsto F^y(w(y)-)$ are increasing.
  (Classical) conformal predictive system

Alternative

Use other in-sample calibrated procedures.

Auto-calibration

Let $(x_1, y_1), \dots, (x_m, y_m) \in \mathcal{X} \times \mathbb{R}$.

- Let $B_1, \dots, B_{m'}$ be a partition of $\{1, \dots, m\}$.
- $$F_{x_k}(y) = \frac{1}{|B_i|} \sum_{j \in B_i} \mathbb{1}\{y_j \le y\}, \quad k \in B_i, y \in \mathbb{R}$$

  is in-sample auto-calibrated, that is,

  $$\hat{\mathbb{P}}_m(Y \le y \mid F_X) = F_X(y), \quad y \in \mathbb{R},$$

  hence, in particular, isotonically calibrated, threshold calibrated, quantile calibrated, and probabilistically calibrated.
  Here, $(X, Y) \sim \hat{\mathbb{P}}_m$, and $\hat{\mathbb{P}}_m$ is the empirical distribution of $(x_j, y_j)_{j=1}^m$.
- We call this a *binning procedure*.
- All in-sample auto-calibrated procedures are of this form.
- Choice: How is the partition constructed?

Auto-calibration

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathbb{R}$.

- Let $B_1, \ldots, B_{m'}$ be a partition of $\{1, \ldots, m\}$.
- $$F_{x_k}(y) = \frac{1}{|B_i|} \sum_{j \in B_i} \mathbb{1}\{y_j \leq y\}, \quad k \in B_i, y \in \mathbb{R}$$

  is in-sample auto-calibrated, that is,

  $$\hat{\mathbb{P}}_m(Y \leq y \mid F_X) = F_X(y), \quad y \in \mathbb{R},$$

  hence, in particular, isotonically calibrated, threshold calibrated, quantile calibrated, and probabilistically calibrated.
  Here, $(X, Y) \sim \hat{\mathbb{P}}_m$, and $\hat{\mathbb{P}}_m$ is the empirical distribution of $(x_j, y_j)_{j=1}^m$.
- We call this a *binning procedure*.
- All in-sample auto-calibrated procedures are of this form.
- Choice: How is the partition constructed?

Auto-calibration

Let $(x_1, y_1), \dots, (x_m, y_m) \in \mathcal{X} \times \mathbb{R}$.

- Let $B_1, \dots, B_{m'}$ be a partition of $\{1, \dots, m\}$.
-
$$F_{x_k}(y) = \frac{1}{|B_i|} \sum_{j \in B_i} \mathbb{1}\{y_j \leq y\}, \quad k \in B_i, y \in \mathbb{R}$$

  is in-sample auto-calibrated, that is,

$$\hat{\mathbb{P}}_m(Y \leq y \mid F_X) = F_X(y), \quad y \in \mathbb{R},$$

  hence, in particular, isotonically calibrated, threshold calibrated, quantile calibrated, and probabilistically calibrated.
  Here, $(X, Y) \sim \hat{\mathbb{P}}_m$, and $\hat{\mathbb{P}}_m$ is the empirical distribution of $(x_j, y_j)_{j=1}^m$.
- We call this a *binning procedure*.
- All in-sample auto-calibrated procedures are of this form.
- Choice: How is the partition constructed?

Auto-calibration

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathbb{R}$.

- Let $B_1, \ldots, B_{m'}$ be a partition of $\{1, \ldots, m\}$.
- $$F_{x_k}(y) = \frac{1}{|B_i|} \sum_{j \in B_i} \mathbb{1}\{y_j \leq y\}, \quad k \in B_i, y \in \mathbb{R}$$

  is in-sample auto-calibrated, that is,

  $$\hat{\mathbb{P}}_m(Y \leq y \mid F_X) = F_X(y), \quad y \in \mathbb{R},$$

  hence, in particular, isotonically calibrated, threshold calibrated, quantile calibrated, and probabilistically calibrated.
  Here, $(X, Y) \sim \hat{\mathbb{P}}_m$, and $\hat{\mathbb{P}}_m$ is the empirical distribution of $(x_j, y_j)_{j=1}^m$.
- We call this a *binning procedure*.
- All in-sample auto-calibrated procedures are of this form.
- Choice: How is the partition constructed?

Let $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{X} \times \mathbb{R}$ be exchangeable.

Let Π be constructed with a binning procedure:

- Let $F^z_{X_k}$ be the binning CDF constructed with $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, z)$.
- Define

$$\Pi_{\ell, X_{n+1}}(y) = \inf\{F^z_{X_{n+1}}(y) \mid z \in \mathbb{R}\}, \quad \Pi_{u, X_{n+1}}(z) = \sup\{F^z_{X_{n+1}}(y) \mid z \in \mathbb{R}\},$$

Theorem (Conformal calibration guarantee)

Predictive system contains an auto-calibrated CDF:

$$F^{Y_{n+1}}_{X_{n+1}}(y) = \mathbb{P}(Y_{n+1} \leq y \mid F^{Y_{n+1}}_{X_{n+1}}), \quad y \in \mathbb{R},$$

and

$$\Pi_{\ell, X_{n+1}}(y) \leq F^{Y_{n+1}}_{X_{n+1}}(y) \leq \Pi_{u, X_{n+1}}(y), \quad y \in \mathbb{R}$$

Isotonic calibration

- Middle ground between probabilistic and auto-calibration
- Based on Isotonic Distributional Regression (IDR) (Henzi, Ziegel, and Gneiting, 2021)

IDR estimator Let $\leq$ be a partial order on $\mathcal{X}$.

Define $\hat{\mathbf{F}} = (F_{x_k})_{k=1}^m$ as

$$\hat{\mathbf{F}} = \underset{F_i \preceq_{\mathrm{st}} F_j \text{ if } x_i \leq x_j}{\operatorname{argmin}} \sum_{\ell=1}^{m} \mathrm{CRPS}(F_\ell, y_\ell).$$

Continuous ranked probability score (CRPS)

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{y \leq z\})^2 \,\mathrm{d}z$$

Why IDR?

- Non-parametric distributional regression procedure under order constraints
- Explicit expression for estimator available
- Implementations available (R and Python)
- Consistency results available (under regularity conditions)

Theorem (In-sample isotonic calibration of IDR)

IDR is in-sample isotonically calibrated, that is,

$$\hat{\mathbb{P}}_m(Y > y \mid \mathcal{A}(F_X^Y)) = 1 - F_X^Y(y), \quad y \in \mathbb{R},$$

and hence, in particular, threshold calibrated, quantile calibrated, and probabilistically calibrated. Here, $(X, Y) \sim \hat{\mathbb{P}}_m$, and $\hat{\mathbb{P}}_m$ is the empirical distribution of $(x_j, y_j)_{j=1}^m$.

Henzi, Ziegel, and Gneiting (2021); Arnold and Ziegel (2025)

- Choice: How is the partial order on $\mathcal{X}$ constructed?

Why IDR?

- Non-parametric distributional regression procedure under order constraints
- Explicit expression for estimator available
- Implementations available (R and Python)
- Consistency results available (under regularity conditions)

Theorem (In-sample isotonic calibration of IDR)

IDR is in-sample isotonically calibrated, that is,

$$\hat{\mathbb{P}}_m(Y > y \mid \mathcal{A}(F_X^Y)) = 1 - F_X^Y(y), \quad y \in \mathbb{R},$$

and hence, in particular, threshold calibrated, quantile calibrated, and probabilistically calibrated. *Here, $(X, Y) \sim \hat{\mathbb{P}}_m$, and $\hat{\mathbb{P}}_m$ is the empirical distribution of $(x_j, y_j)_{j=1}^m$.*

Henzi, Ziegel, and Gneiting (2021); Arnold and Ziegel (2025)

- Choice: How is the partial order on $\mathcal{X}$ constructed?

Why IDR?

- Non-parametric distributional regression procedure under order constraints
- Explicit expression for estimator available
- Implementations available (R and Python)
- Consistency results available (under regularity conditions)

Theorem (In-sample isotonic calibration of IDR)

IDR is in-sample isotonically calibrated, that is,

$$\hat{\mathbb{P}}_m(Y > y \mid \mathcal{A}(F_X^Y)) = 1 - F_X^Y(y), \quad y \in \mathbb{R},$$

and hence, in particular, threshold calibrated, quantile calibrated, and probabilistically calibrated. Here, $(X, Y) \sim \hat{\mathbb{P}}_m$, and $\hat{\mathbb{P}}_m$ is the empirical distribution of $(x_j, y_j)_{j=1}^m$.

Henzi, Ziegel, and Gneiting (2021); Arnold and Ziegel (2025)

- Choice: How is the partial order on $\mathcal{X}$ constructed?

Let $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \in \mathcal{X} \times \mathbb{R}$ be exchangeable.

Let Π be constructed with IDR (*conformal IDR*):

- Let $F^z_{X_k}$ be the IDR CDF computed from $(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, z)$.
- Define

$$\Pi_{\ell, X_{n+1}}(y) = \inf\{F^z_{X_{n+1}}(y) \mid z \in \mathbb{R}\}, \quad \Pi_{u, X_{n+1}}(z) = \sup\{F^z_{X_{n+1}}(y) \mid z \in \mathbb{R}\},$$

Theorem (Conformal calibration guarantee)

Predictive system contains an isotonically calibrated CDF:

$$F^{Y_{n+1}}_{X_{n+1}}(y) = 1 - \mathbb{P}(Y_{n+1} > y \mid \mathcal{A}(F^{Y_{n+1}}_{X_{n+1}})), \quad y \in \mathbb{R},$$

and

$$\Pi_{\ell, X_{n+1}}(y) \leq F^{Y_{n+1}}_{X_{n+1}}(y) \leq \Pi_{u, X_{n+1}}(y), \quad y \in \mathbb{R}$$

Comments

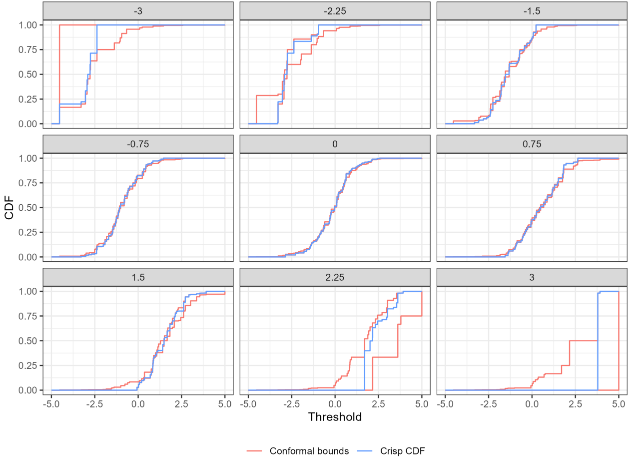
- Conformal guarantee does not depend of any isotonicity assumption.
- The partial order on $\mathcal{X}$ can be estimated on the same sample (computational challenge! "full conformal") or on an independent sample ("split conformal").

Thickness of predictive systems

- Predictive systems are only useful if they are thin.
- Classical conformal predictive systems:
  - Thickness is $1/(n+1)$.
- Auto-calibration: Binning procedures, where bins are determined only based on $X_1, \ldots, X_{n+1}$ (example: k-means clustering):
  - Thickness is $1/(\text{size of bin containing } n+1)$.
- Isotonic calibration with IDR:
  - Expected thickness is less or equal to $14n^{-1/6}$.

Tiny simulation example for conformal IDR

$$X \sim \mathcal{N}(0,1),\ Y \sim \mathcal{N}(X,1),\ n = 512.$$



- Principled approach to choose a crisp conformal IDR.
- Expected thickness goes to zero asymptotically.
- Thickness of conformal IDR informs about epistemic uncertainty.

Aleatoric and epistemic uncertainty

Aleatoric uncertainty

Aleatoric uncertainty of future outcome Y is fully described by

$$\mathcal{L}(Y \mid X).$$

Uncertainty remains even with infinite amounts of data (X_i, Y_i).

Epistemic uncertainty (second order probabilities, ambiguity, ...)

Uncertainty due to our approximation of $\mathcal{L}(Y \mid X)$ based on limited data, limited knowledge of data generating process, parameter estimation, ...
Uncertainty goes away if we have infinite amounts of data.

- With IDR we recover $\mathcal{L}(Y \mid \mathcal{A}(X))$.

Aleatoric and epistemic uncertainty

Aleatoric uncertainty

Aleatoric uncertainty of future outcome Y is fully described by

$$\mathcal{L}(Y \mid X).$$

Uncertainty remains even with infinite amounts of data (X_i, Y_i).

Epistemic uncertainty (second order probabilities, ambiguity, ...)

Uncertainty due to our approximation of $\mathcal{L}(Y \mid X)$ based on limited data, limited knowledge of data generating process, parameter estimation, ...
Uncertainty goes away if we have infinite amounts of data.

- With IDR we recover $\mathcal{L}(Y \mid \mathcal{A}(X))$.

Aleatoric and epistemic uncertainty

Aleatoric uncertainty

Aleatoric uncertainty of future outcome Y is fully described by

$$\mathcal{L}(Y \mid X).$$

Uncertainty remains even with infinite amounts of data (X_i, Y_i).

Epistemic uncertainty (second order probabilities, ambiguity, ...)

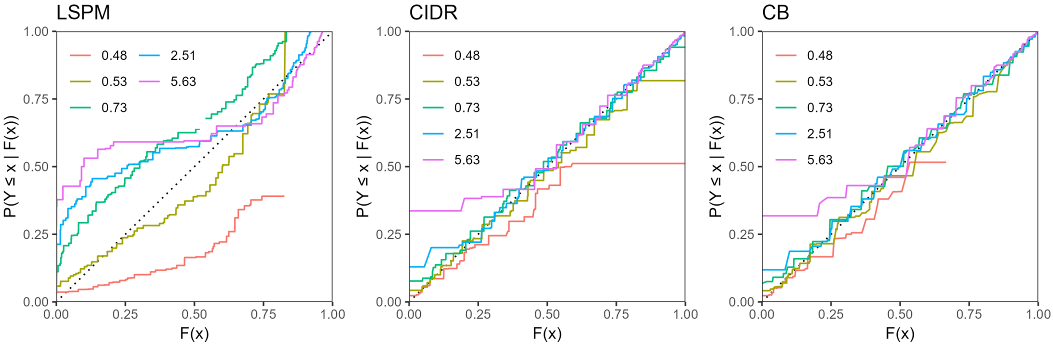
Uncertainty due to our approximation of $\mathcal{L}(Y \mid X)$ based on limited data, limited knowledge of data generating process, parameter estimation, ...
Uncertainty goes away if we have infinite amounts of data.

- With IDR we recover $\mathcal{L}(Y \mid \mathcal{A}(X))$.

Case study: Length of stay in intensive care units

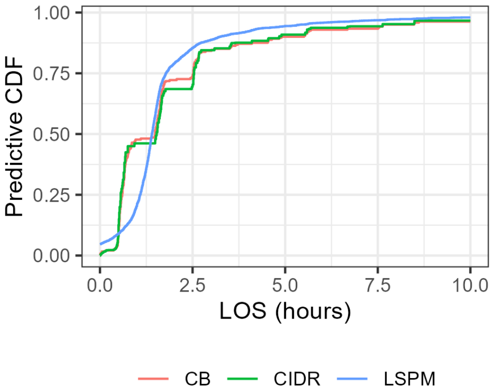
- Predictions for individual patients' length of stay in ICU's in Switzerland 24h after admission[1]

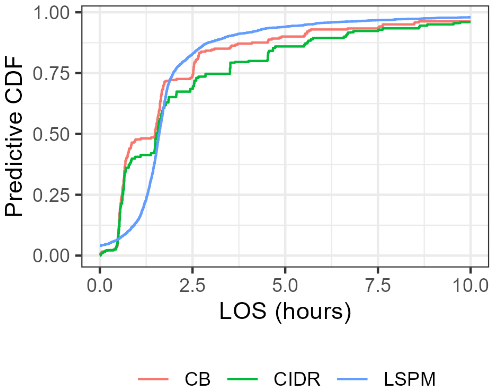
Threshold calibration



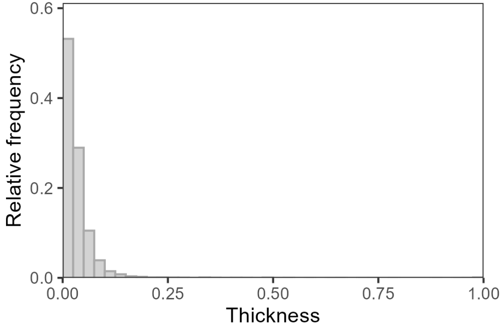
[1] Data provided by *G.-R. Kleger and Schweizerische Gesellschaft für Intensivmedizin*. Data is internal hospital data and not publicly available.

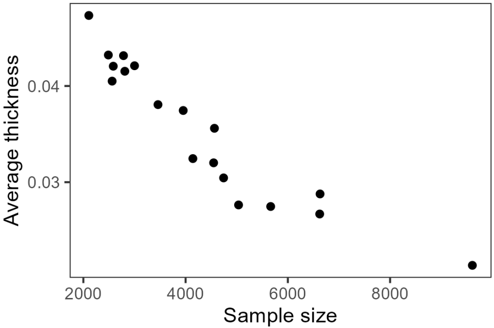
Examples of predictive cdfs





Epistemic uncertainty assessment with conformal IDR





Summary

- In-sample calibration yields conformal calibration guarantees.
- Strong out-of-sample calibration guarantees are possible.
- Arguments can be extended to distribution shifts.
- Conformal binning is simple but works well.
  Only example explored so far: k-means clustering.
- Conformal IDR allows to quantify epistemic uncertainty, since IDR converges to a well-understood limiting object.
- Outlook: Conformal calibration guarantees for point predictions.

Outlook: Conformal calibration guarantees for point predictions

- $Y \in \mathbb{R}$.
  - Claim size. ($Y \in [0, \infty) \subseteq \mathbb{R}$)
- Point prediction for Y:
  - Single valued "best guess" $Z \in \mathcal{Y}$.
  - Does not quantify uncertainty, but maybe useful/necessary e.g. for pricing.
  - If X is information available for prediction, often, Z should approximate $\mathbb{E}[Y \mid X]$.

Definition

A prediction $Z \in \mathbb{R}$ for $Y \in \mathbb{R}$ is *expectation-calibrated* if

$$\mathbb{E}[Y \mid Z] = Z.$$

Conformal calibration guarantee:

Construct (a small) set $\mathcal{C}_{n+1}$ such that

$$\mathbb{P}\big(\mathbb{E}[Y_{n+1} \mid Z_{n+1}] \in \mathcal{C}_{n+1}\big) \geq 1 - \alpha.$$

Outlook: Conformal calibration guarantees for point predictions

- $Y \in \mathbb{R}$.
  - Claim size. ($Y \in [0, \infty) \subseteq \mathbb{R}$)
- Point prediction for Y:
  - Single valued "best guess" $Z \in \mathcal{Y}$.
  - Does not quantify uncertainty, but maybe useful/necessary e.g. for pricing.
  - If X is information available for prediction, often, Z should approximate $\mathbb{E}[Y \mid X]$.

Definition

A prediction $Z \in \mathbb{R}$ for $Y \in \mathbb{R}$ is *expectation-calibrated* if

$$\mathbb{E}[Y \mid Z] = Z.$$

Conformal calibration guarantee:

Construct (a small) set $\mathcal{C}_{n+1}$ such that

$$\mathbb{P}\big(\mathbb{E}[Y_{n+1} \mid Z_{n+1}] \in \mathcal{C}_{n+1}\big) \geq 1 - \alpha.$$

Outlook: Conformal calibration guarantees for point predictions

- $Y \in \mathbb{R}$.
  - Claim size. ($Y \in [0, \infty) \subseteq \mathbb{R}$)
- Point prediction for Y:
  - Single valued "best guess" $Z \in \mathcal{Y}$.
  - Does not quantify uncertainty, but maybe useful/necessary e.g. for pricing.
  - If X is information available for prediction, often, Z should approximate $\mathbb{E}[Y \mid X]$.

Definition

A prediction $Z \in \mathbb{R}$ for $Y \in \mathbb{R}$ is *expectation-calibrated* if

$$\mathbb{E}[Y \mid Z] = Z.$$

Conformal calibration guarantee:

Construct (a small) set $\mathcal{C}_{n+1}$ such that

$$\mathbb{P}\big(\mathbb{E}[Y_{n+1} \mid Z_{n+1}] \in \mathcal{C}_{n+1}\big) \geq 1 - \alpha.$$

References

S. Allen, G. Gavrilopoulos, A. Henzi, G.-R. Kleger, and J. Ziegel. In-sample calibration yields conformal calibration guarantees. *Preprint, arXiv:2503.03841*, 2025a.

S. Allen, J. Koh, J. Segers, and J. Ziegel. Tail calibration of probabilistic forecasts. *Journal of the American Statistical Association*, 2025b. To appear.

S. Arnold and J. Ziegel. Isotonic conditional laws. *Bernoulli*, 31:1140–1159, 2025.

T. Dimitriadis, T. Gneiting, and A. I. Jordan. Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118: e2016191118, 2021.

A. Henzi, J. F. Ziegel, and T. Gneiting. Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B*, 85:963–993, 2021.

R. Ranjan and T. Gneiting. Combining probability forecasts. *Journal of the Royal Statistical Society: Series B*, 72:71–91, 2010.

Thank you!

Why the CRPS?

It is a strictly proper scoring rule.

If $Y \sim F$ and G is any other CDF, then $S(F, y)$ is *strictly proper* if

$$\mathbb{E}_F S(F, Y) \leq \mathbb{E}_F S(G, Y)$$

with equality if and only if $F = G$.

Example 1

If F, G have finite mean, then the CRPS

$$\text{CRPS}(F, Y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{Y \leq z\})^2 \, \mathrm{d}z$$

is strictly proper.

Example 2

If F, G have densities f, g, then the logarithmic score

$$S_{\log}(F, y) = -\log f(y)$$

is strictly proper.

Why the CRPS?

It is a strictly proper scoring rule.

If $Y \sim F$ and G is any other CDF, then $S(F, y)$ is *strictly proper* if

$$\mathbb{E}_F S(F, Y) \leq \mathbb{E}_F S(G, Y)$$

with equality if and only if $F = G$.

Example 1

If F, G have finite mean, then the CRPS

$$\mathrm{CRPS}(F, Y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{Y \leq z\})^2 \, \mathrm{d}z$$

is strictly proper.

Example 2

If F, G have densities f, g, then the logarithmic score

$$S_{\log}(F, y) = -\log f(y)$$

is strictly proper.

Why the CRPS?

It is a strictly proper scoring rule.

If $Y \sim F$ and G is any other CDF, then $S(F, y)$ is *strictly proper* if

$$\mathbb{E}_F S(F, Y) \leq \mathbb{E}_F S(G, Y)$$

with equality if and only if $F = G$.

Example 1

If F, G have finite mean, then the CRPS

$$\text{CRPS}(F, Y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{Y \leq z\})^2 \, \mathrm{d}z$$

is strictly proper.

Example 2

If F, G have densities f, g, then the logarithmic score

$$S_{\log}(F, y) = -\log f(y)$$

is strictly proper.

Mathematical setup

"If the covariate increases we expect an increase of the outcome."

$$x \leq x' \implies \mathcal{L}(Y \mid X = x) \preceq_{\text{st}} \mathcal{L}(Y \mid X = x')$$

$$\iff F_{Y|X=x}(y) \geq F_{Y|X=x'}(y), \quad y \in \mathbb{R}$$

$$\iff q_\alpha(Y|X = x) \leq q_\alpha(Y|X = x'), \quad \alpha \in (0,1)$$

IDR estimator (for $x \in \mathbb{R}$): Data $(x_i, y_i)_{i=1}^n$, $x_1 < \cdots < x_n$

Define $\hat{\mathbf{F}} = (\hat{F}_i)_{i=1}^n = (\hat{F}_{Y|X=x_i})_{i=1}^n$ as

$$\hat{\mathbf{F}} = \underset{F_1 \preceq_{\text{st}} \cdots \preceq_{\text{st}} F_n}{\operatorname{argmin}} \sum_{\ell=1}^{n} \text{CRPS}(F_\ell, y_\ell).$$

Continuous ranked probability score (CRPS)

$$\text{CRPS}(F, Y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{Y \leq z\})^2 \, \mathrm{d}z$$

Mathematical setup

"If the covariate increases we expect an increase of the outcome."

$$x \leq x' \implies \mathcal{L}(Y \mid X = x) \preceq_{\text{st}} \mathcal{L}(Y \mid X = x')$$

$$\iff F_{Y|X=x}(y) \geq F_{Y|X=x'}(y), \quad y \in \mathbb{R}$$

$$\iff q_\alpha(Y|X = x) \leq q_\alpha(Y|X = x'), \quad \alpha \in (0, 1)$$

IDR estimator (for $x \in \mathbb{R}$): Data $(x_i, y_i)_{i=1}^n$, $x_1 < \cdots < x_n$

Define $\hat{\mathbf{F}} = (\hat{F}_i)_{i=1}^n = (\hat{F}_{Y|X=x_i})_{i=1}^n$ as

$$\hat{\mathbf{F}} = \underset{F_1 \preceq_{\text{st}} \cdots \preceq_{\text{st}} F_n}{\operatorname{argmin}} \sum_{\ell=1}^{n} \text{CRPS}(F_\ell, y_\ell).$$

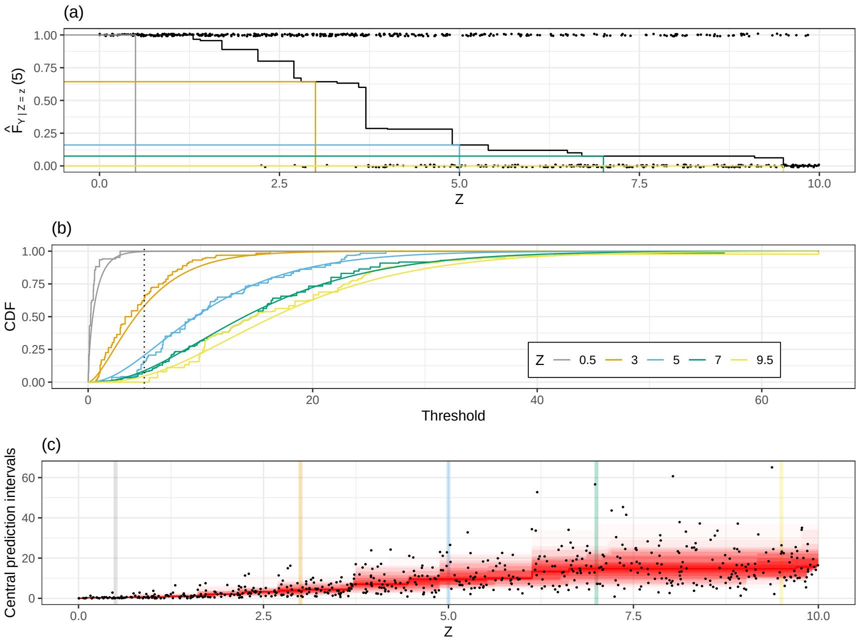
Continuous ranked probability score (CRPS)

$$\text{CRPS}(F, Y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{Y \leq z\})^2 \, dz$$

Then

$$\hat{F}_{Y|X=x_i} = \hat{F}_i(y) \;=\; \max_{j=i,\dots,n} \min_{k=1,\dots,j} \frac{1}{j-k+1} \sum_{\ell=k}^{j} \mathbb{1}\{y_\ell \leq y\}.$$

- $\hat{F}_1(y), \dots, \hat{F}_n(y)$ is the antitonic regression of the binary outcomes $\mathbb{1}\{y_1 \leq y\}, \dots, \mathbb{1}\{y_n \leq y\}$.



(a)
$\hat{F}_{Y|Z=z}(5)$
1.00
0.75
0.50
0.25
0.00
0.0
2.5
5.0
7.5
10.0
Z
(b)
CDF
Threshold
0
20
40
60
Z
0.5
3
5
7
9.5
(c)
Central prediction intervals
0
20
40
60
Z